Psi-BLAST
User Manual

Table des matières

Introduction

PSI-BLAST (Position-Specific Iterative BLAST) is a protein BLAST search that uses a PSSM (position-specific scoring matrix) as a query instead of an individual sequence.

PSI-BLAST refers to a feature of BLAST 2.0 in which a profile (or position specific scoring matrix, PSSM) is constructed (automatically) from a multiple alignment of the highest scoring hits in an initial BLAST search. The PSSM is generated by calculating position-specific scores for each position in the alignment. Highly conserved positions receive high scores and weakly conserved positions receive scores near zero. The profile is used to perform a second (etc.) BLAST search and the results of each "iteration" used to refine the profile. This iterative searching strategy results in increased sensitivity.

Reference : Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410

Parameters

Proteic Databank

Select a sequence databank versus which theresearch will be launched :

- **NR** : Non-redundant Proteicsequences collection maintained at NCBI.
- **Uniprot** (Universal Protein Resource) : Proteic sequences collection merging informations from 3 databanks : Swiss-Prot, TrEMBL, and PIR.
- **PIR** (Protein Information Ressource) : Proteic sequences collection, maintained by Georgetown University Medical Center (USA).
- **SwissProt**: Non-redundant Proteic sequences collection (produced by sequencing) maintained by Expasy (Switzerland).
- **Genpept** : GenPept is produced by parsing the corresponding GenBank (NCBI) release for translated coding regions as definedin the GenBank FEATURES section of each sequence.
- **RefSeq Protein** : The RefSeq (Reference Sequence) collection provide a comprehensive, integrated, non-redundant set of sequences, including genomic DNA, transcript (RNA), and protein products. This databank is managed by NCBI.
- **PDB** (Protein Data Bank) : protein with known 3D structure sequences.

Expect

This setting specifies the statistical significance threshold for reporting matches against database sequences. The default value (10) means that 10 such matches are expected to be found merely by chance, according to the stochastic model of Karlin and Altschul (1990). If the statistical significance ascribed to a match is greater than the EXPECT threshold, the match will not be reported. Lower EXPECT thresholds are more stringent, leading to fewer chance matches being reported.

Expect (inclusion)

This parameter has the same statistical meaning as explained before. The value given here is the threshold to report a match from one iteration to the next iteration.

Matrix

A key element in evaluating the quality of a pairwise sequence alignment is the "substitution matrix", which assigns a score for aligning any possible pair of residues. The matrix used in a BLAST search can be changed depending on the type of sequences you are searching with (see the BLAST Frequently Asked Questions).

Substitution Matrix :

A substitution matrix containing values proportional to the probability that amino acid i mutates into amino acid j for all pairs of amino acids.

The PAM family :

- PAM matrices are based on global alignments of closely related proteins.
- The PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence.
- Other PAM matrices are extrapolated from PAM1.

The BLOSUM family :

- BLOSUM matrices are based on local alignments.
- BLOSUM 62 is a matrix calculated from comparisons of sequences with no less than 62% divergence.
- All BLOSUM matrices are based on observed alignments; they are not extrapolated from comparisons of closely related proteins.
- BLOSUM 62 is the default matrix in BLAST 2.0. Though it is tailored for comparisons of moderately distant proteins, it performs well in detecting closer relationships. A search for distant relatives may be more sensitive with a different matrix.

Note : If you're not sure about your matrix choice, use BLOSUM62.

SEG Filter

This function mask off segments of the query sequence that have low compositional complexity, as determined by the SEG program of Wootton and Federhen (Computers and Chemistry, 1993). Filtering can eliminate statistically significant but biologically uninteresting reports from the blast output (e.g., hits against common acidic-, basic- or proline-rich regions), leaving the more biologically interesting regions of the query sequence available for specific matching against database sequences.

Filtering is only applied to the query sequence (or its translation products), not to database sequences. Default filtering is SEG for psi-BLAST program.

SEG : A program for filtering low complexity regions in amino acid sequences. Residues that have been masked are represented as "X" in an alignment.

Maximum number of passes

It is the maximum number of passes to use during the Psi-blast's research in multipass.

Gap costs

Increasing the Gap Costs and Lambda ratio will result in alignments which decrease the number of Gaps introduced.